W412–W416 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 27 May 2010*
doi:10.1093/nar/gkq474

PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences

Yoichi Murakami[1], Ruth V. Spriggs[2], Haruki Nakamura[1] and Susan Jones[2,*]

[1]Laboratory of Protein Informatics, Research Center for Structural and Functional Proteomics, Institute for Protein Research, Osaka University, Osaka, Japan and [2]Department of Chemistry and Biochemistry, School of Life Sciences, John Maynard-Smith Building, University of Sussex, Falmer BN1 9QG, UK

Received January 30, 2010; Revised May 6, 2010; Accepted May 13, 2010

ABSTRACT

The PiRaNhA web server is a publicly available online resource that automatically predicts the location of RNA-binding residues (RBRs) in protein sequences. The goal of functional annotation of sequences in the field of RNA binding is to provide predictions of high accuracy that require only small numbers of targeted mutations for verification. The PiRaNhA server uses a support vector machine (SVM), with position-specific scoring matrices, residue interface propensity, predicted residue accessibility and residue hydrophobicity as features. The server allows the submission of up to 10 protein sequences, and the predictions for each sequence are provided on a web page and via email. The prediction results are provided in sequence format with predicted RBRs highlighted, in text format with the SVM threshold score indicated and as a graph which enables users to quickly identify those residues above any specific SVM threshold. The graph effectively enables the increase or decrease of the false positive rate. When tested on a non-redundant data set of 42 protein sequences not used in training, the PiRaNhA server achieved an accuracy of 85%, specificity of 90% and a Matthews correlation coefficient of 0.41 and outperformed other publicly available servers. The PiRaNhA prediction server is freely available at http://www.bioinformatics.sussex.ac.uk/PIRANHA.

INTRODUCTION

RNA-binding proteins (RBPs) play key roles in many cellular processes, including gene expression regulation. The fundamental role of RBPs within the cell is reflected in the wide range of human diseases, including neurological disorders and cancer, to which they have been linked (1). However, proteomic sequence data includes a significant percentage of RBPs in which the location of RNA-binding residues (RBRs) is unknown. Hence, computational methods to predict the location of the RBRs in such proteins are of great significance.

The availability of the structures of an increasing number of protein–RNA complexes has allowed RBRs to be characterized using features such as hydrophobicity, solvent accessibility, evolutionary conservation and charge (2–4). Such analysis of RBRs has led to the development of methods to predict RBRs from protein sequence information, using machine learning techniques (5–10). The identification of RNA-binding sites from protein sequence information alone is critically important for understanding the function of RBPs when the structure of the protein–RNA complex is not known. Three of the prediction methods are currently available as publicly available web servers: RNABindR (6), BindN (7) and PPRInt (9). The majority of these methods use inclusive definitions of RBRs, based on lenient distance constraints between the RNA and the protein entities (between 5 Å and 6 Å). These methods also use either position-specific scoring matrices (PSSMs) (to estimate evolutionary conservation) or residue parameters as features to differentiate RBRs from non-RBRs. These two separate approaches give results with varying levels of accuracy, and combined with their inclusive definition of RBRs can give large numbers of false positive predictions (11).

This article presents a new publicly available online server (**PiRaNhA**, capitals denote **P**rotein-**RNA**; pronounced ‘piranha’) for the prediction of RBRs from protein sequence information alone. This server differs from other prediction servers in three important criteria: (i) the server is based on a support vector machine (SVM) model that has been trained on known structures of protein–RNA complexes from the Protein Data Bank

*To whom correspondence should be addressed. Tel: +44 0 1273 877553; Fax: +44 1273 678297; Email: s.jones@sussex.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

(PDB) (12). The RBRs in these structures are defined using a restrictive distance constraint based on intermolecular interaction data calculated using the HBPLUS software (13). RBRs are defined as those residues making an intermolecular hydrogen bond or van der Waals contact, with a distance of ≤3.9 Å. (ii) The SVM model integrates PSSMs with three physicochemical residue parameters to identify RBRs. The three parameters are residue interface propensity, predicted residue accessibility and residue hydrophobicity. (iii) The server provides the unique facility for users to make predictions for multiple sequences. The PiRaNhA server achieves a Matthews correlation coefficient (MCC, a balanced performance measure that includes the numbers of true and false, positive and negative, predictions) of 0.41 for a data set of 42 known RNA-binding protein sequences not used in training. Thus, PiRaNhA outperforms other web server prediction tools; RNABindR (6), BindN (7) and PPRInt (9), which achieve an MCC of 0.36, 0.29 and 0.34, respectively, on the same data set (11). The PiRaNhA server allows researchers to upload the sequence of one or more proteins and obtain a set of RBR predictions. The accurate nature of the predictions means fewer site-directed mutagenesis experiments are required to verify the RNA-binding function of residues. The PiRaNhA web server is freely available at http://www.bioinformatics.sussex.ac.uk/PIRANHA.

METHODS

Server input

The PiRaNhA server is designed to be easy to use and to provide results that are easily interpreted. It allows the submission of single or multiple (up to a maximum of 10) protein sequences in FASTA format, by cutting and pasting to the submission page or by file upload. Protein sequences of unlimited length are accepted, but the calculation time required for the PSI-BLAST alignments required for the PSSM generation is related to the sequence length. Hence, for sequences longer than 150 residues selection of the email option is recommended (as is the case with batch submissions), so results can be emailed when the calculations are complete.

Sequence feature vectors

For all residues in the submitted protein sequence(s), four sequence features are calculated and encoded into feature vectors in the SVM model. These features are: **(i)** a PSSM created using PSI-BLAST (14) with an *E*-value threshold of 0.001 for three iterations, and the NCBI nr sequence database. This feature describes the evolutionary conservation of the residue positions. **(ii)** A residue interface propensity that describes the likelihood of a residue of a specific type being found in an RNA-binding site. The interface propensity values are taken from our previous analysis of known RNA-binding sites (3). **(iii)** A predicted residue accessibility value that quantifies the estimated solvent exposure of each residue and is calculated using SABLE (15). **(iv)** A residue hydrophobicity score based on the Kyte and Doolittle hydropathy scale (16). These four properties are integrated into a single feature vector based on a sequence window of 23 residues, with the residue being described in the centre. The calculation of the PSSMs and the prediction of accessible surface area are computationally intensive, and it is for this reason that users may want to obtain results via email when making a batch submission of more than one protein sequence.

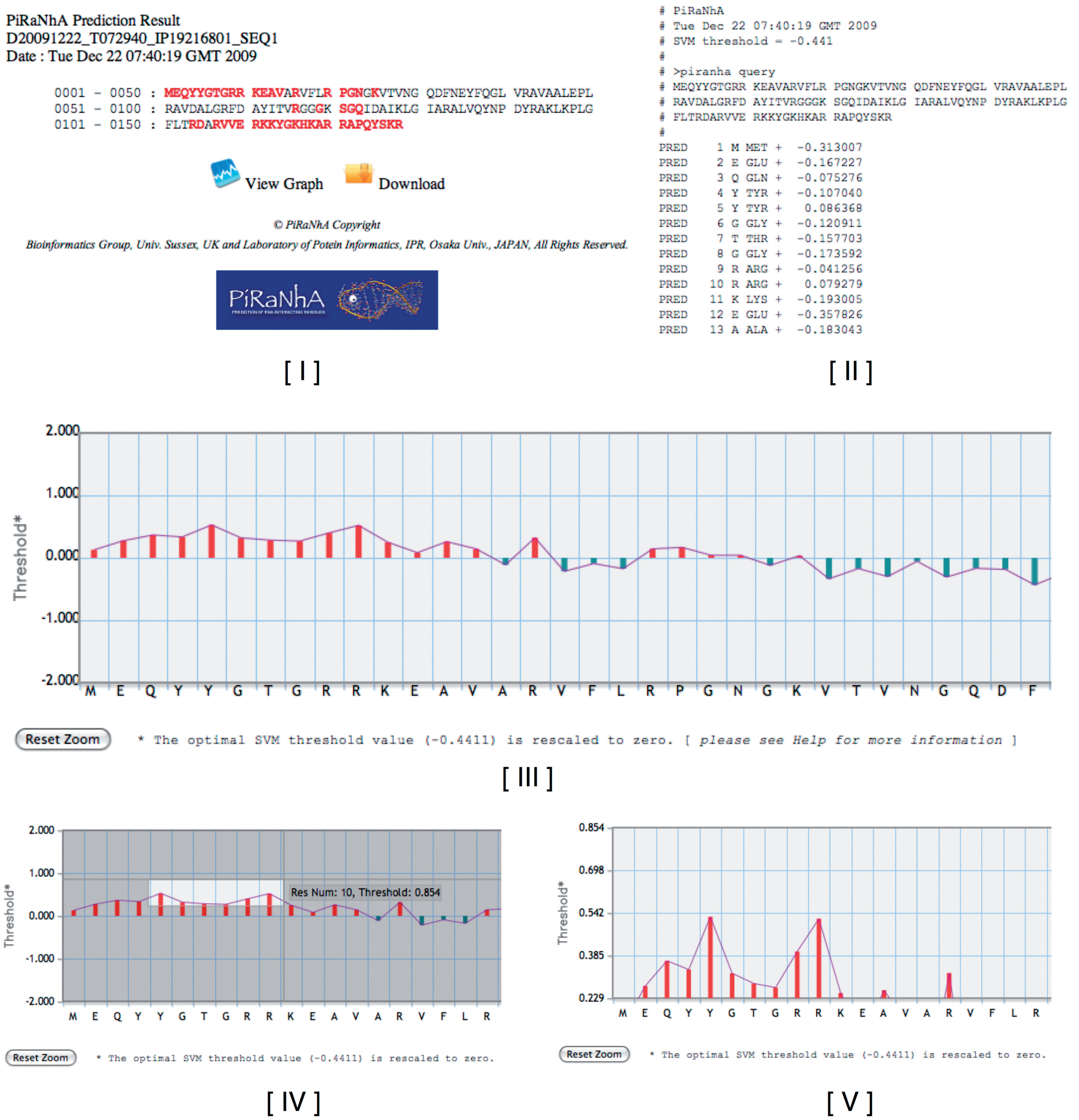
The SVM model

The SVM model used in the PiRaNhA server is based on LIBSVM (Chang C-C and Lin C-J: LIBSVM: a library for SVMs, 2001. www.csie.ntu.edu.tw/~cjlin/libsvm, version 2.8) and uses the Radial Basis Function kernel. In order to distinguish between RBRs and non-RBRs in protein sequences, the SVM model was trained on a non-redundant set of 81 known RNA-binding protein sequences (RNAset81) whose complexed structures are known. The model was tested on a non-redundant data set of 42 known RNA-binding protein sequences not used in training (RNAtestset42); of the 8554 residues in these sequences, 14.8% were known RBRs. The SVM model achieved an MCC of 0.41, a sensitivity of 53%, a specificity of 90% and a precision of 48% and outperforms the other publicly available servers [RNABindR (6), BindN (7) and PPRInt (9)] tested on the same data set (11).

RESULTS

The PiRaNhA server provides predictions for each submitted sequence in a separate web page and, if requested, by email with links to URLs and attached files. The results are provided in three formats: (i) the user's submitted sequence shown with residues involved in protein–RNA interactions highlighted in red (Figure 1I), (ii) a text file of raw prediction results with SVM scores that can be downloaded (Figure 1II) and (iii) a graph where the submitted sequence is plotted against SVM threshold values (Figure 1III). On the graph, the *x*-axis shows the submitted sequence and the *y*-axis shows the threshold for the prediction. The optimal threshold value (–0.4411) [which achieves an MCC of 0.50, and an area under ROC curve (AUC) of 0.86 in a 5-fold cross-validation on the training data set (11)] is rescaled to zero for the ease of interpretation. The graph has a built in 'click and drag' zoom function to enable users to highlight and select residues above a desired threshold (Figure 1IV and V). By increasing the threshold value, a user can effectively decrease the false positive rate of predictions. This graph function enables users to quickly determine which residues are most likely to bind to RNA, and hence be the initial targets for site-directed mutagenesis.

Figure 2 shows examples of two PiRaNhA predictions taken from the RNAtestset42 data set; (Figure 2A) 30S ribosomal protein S9 (PDB-ID 2J00, chain I), (Figure 2B) CCA-adding enzyme (PDB-ID 2DRB, chain A), in which the overlap between the predictions and the known RBRs from the protein–RNA complex is detailed. This clearly shows the accuracy of the PiRaNhA server, but also



PiRaNhA Prediction Result
D20091222_T072940_IP19216801_SEQ1
Date : Tue Dec 22 07:40:19 GMT 2009

```
0001 - 0050 : MEQYYGTGRR KEAVARVFLR PGNGKVTVNG QDFNEYFQGL VRAVAALEPL
0051 - 0100 : RAVDALGRFD AYITVRGGGK SGQIDAIKLG IARALVQYNP DYRAKLKPLG
0101 - 0150 : FLTRDARVVE RKKYGKHKAR RAPQYSKR
```

© PiRaNhA Copyright

Bioinformatics Group, Univ. Sussex, UK and Laboratory of Potein Informatics, IPR, Osaka Univ., JAPAN, All Rights Reserved.

```
# PiRaNhA
# Tue Dec 22 07:40:19 GMT 2009
# SVM threshold = -0.441
#
# >piranha query
# MEQYYGTGRR KEAVARVFLR PGNGKVTVNG QDFNEYFQGL VRAVAALEPL
# RAVDALGRFD AYITVRGGGK SGQIDAIKLG IARALVQYNP DYRAKLKPLG
# FLTRDARVVE RKKYGKHKAR RAPQYSKR
#
PRED    1 M MET +  -0.313007
PRED    2 E GLU +  -0.167227
PRED    3 Q GLN +  -0.075276
PRED    4 Y TYR +  -0.107040
PRED    5 Y TYR +   0.086368
PRED    6 G GLY +  -0.120911
PRED    7 T THR +  -0.157703
PRED    8 G GLY +  -0.173592
PRED    9 R ARG +  -0.041256
PRED   10 R ARG +   0.079279
PRED   11 K LYS +  -0.193005
PRED   12 E GLU +  -0.357826
PRED   13 A ALA +  -0.183043
```

* The optimal SVM threshold value (-0.4411) is rescaled to zero. [please see Help for more information]

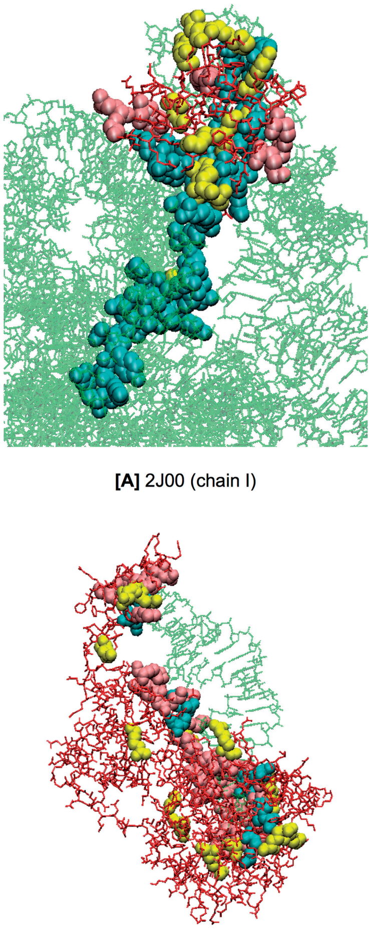
* The optimal SVM threshold value (-0.4411) is rescaled to zero.

* The optimal SVM threshold value (-0.4411) is rescaled to zero.

View Graph
Download
PiRaNhA
PREDICTION OF RNA-INTERACTING RESIDUES
[I]
[II]
2.000
1.000
0.000
-1.000
-2.000
Threshold*
M E Q Y Y G T G R R K E A V A R V F L R P G N G K V T V N G Q D F
Reset Zoom
[III]
2.000
1.000
0.000
-1.000
-2.000
Threshold*
Res Num: 10, Threshold: 0.854
M E Q Y Y G T G R R K E A V A R V F L R
Reset Zoom
[IV]
0.854
0.698
0.542
0.385
0.229
Threshold*
M E Q Y Y G T G R R K E A V A R V F L R
Reset Zoom
[V]

Figure 1. Example PiRaNhA server prediction results for 30S ribosomal protein S9 (PDB-ID 2J00, chain I). (**I**) The sequence format webpage where the predicted RBRs are highlighted in red. (**II**) The text format results, which includes the sequence and the SVM values that can be downloaded. (**III**) The graphical interpretation of the results in which the submitted sequence is plotted against SVM threshold values. The *x*-axis shows the submitted sequence and the *y*-axis the threshold for the prediction. The optimal SVM threshold value (0.4411) is rescaled to zero, for ease of interpretation. The graph has a 'click and drag' zoom function to enable the easy highlighting of residues (**IV**) above a desired threshold to produce a finer grained graph (**V**).

highlights the fact that predictions can include false positives (FPs), i.e. residues predicted to be RNA binding that, when compared to the RNA–protein structure in the PDB, are not defined in the known RNA-binding site. The relatively high number of FPs for some predictions is reflected in a mean precision value of 48%. However, to address this issue the PiRaNhA server allows the user to increase the SVM threshold value above the default (–0.4411). Increasing the threshold value decreases the number of false positives; a strategy that would be



[A] 2J00 (chain I)

[B] 2DRB (chain A)

[A] 2J00 (chain I); 30S Ribosomal Protein S9
TP=42, FP=14, TN=65, FN=6, Sn=87.5%, Sp=82.3%, Acc=84.3%, MCC=0.681, Precision=75.0%

```
001-070    MEQYYGTGRR KEAVARVFLR PGNGKVTVNG QDFNEYFQGL VRAVAALEPL RAVDALGRFD AYITVRGGGK
Interface  ?++-+-+-+- ++-+-+-+-+ ---------- -----+-++- ---------- ---------- -+---+++++
PiRaNhA    ++++++++++ ++++-+-+-+ +++-+---+- ---------- ---------- ---------- -----+++++

071-128    SGQIDAIKLG IARALVQYNP DYRAKLKPLG FLTRDARVVE RKKYGKHKAR RAPQYSKR
Interface  +++------- --+------- --+---+--- ---+--++++ ++++++++-+ ++++++++
PiRaNhA    +++----+-+ --+------- ---------- ---++-++++ ++++++++++ ++++++++
```

[B] 2DRB (chain A); CCA-adding enzyme
TP=10, FP=13, TN=380, FN=34, Sn=22.7%, Sp=96.7%, Acc=89.2%, MCC=0.262, Precision=43.5%

```
001-070    MKVEEILEKA LELVIPDEEE VRKGREAEEE LRRRLDELGV EYVFVGSYAR NTWLKGSLEI DVFLLFPEEF
Interface  ---------- ---------- ---------- ---------- -----+---+ --------+- +-+-------
PiRaNhA    ---------- ---------- ---------- ---------- ------+--+ ---------- ----------

071-140    SKEELRERGL EIGKAVLDSY EIRYAEHPYV HGVVKGVEVD VVPCYKLKEP KNIKSAVDRT PFHHKWLEGR
Interface  ---------- ---------- ---++++-+- ---------+ -+-------- -----++--+ --+-------
PiRaNhA    ---------- ---------- ------+--- ---------- ---------- --------++ --+-+----+

141-210    IKGKENEVRL LKGFLKANGI YGAEYKVRGF SGYLCELLIV FYGSFLETVK NARRWTRRTV IDVAKGEVRK
Interface  ---------- ---------- --+++----- ---------- ---------- ---------- ----------
PiRaNhA    --------+- -----+--+- ---------- ---------- ---------- -------+-- ----------

211-280    GEEFFVVDPV DEKRNVAANL SLDNLARFVH LCREFMEAPS LGFFKPKHPL EIEPERLRKI VEERGTAVFA
Interface  ---------- ---+---++- ---------- ---------- ---------- ---------- ----------
PiRaNhA    ---------- ---+------ ---------- ---------- ---------- ---------- ----------

281-350    VKFRKPDIVD DNLYPQLERA SRKIFEFLER ENFMPLRSAF KASEEFCYLL FECQIKEISR VFRRMGPQFE
Interface  ---------- ++--++--+- --+------+ ---------- ---------- ---------- ---++++---
PiRaNhA    ---------- --------+- ---------- ------+--- ---------- ---------- ---+-+----

351-420    DERNVKKFLS RNRAFRPFIE NGRWWAFEMR KFTTPEEGVR SYASTHWHTL GKNVGESIRE YFEIISGEKL
Interface  ---+--++-- +-+------- ---------- ---------- -+---+-++- ++-------- ----------
PiRaNhA    ---------- ---------- +-+------+ ---------- --------+- +-+------- ----------

421-437    FKEPVTAELC EMMGVKD
Interface  ---------- -------
PiRaNhA    ---------- -------
```

Figure 2. Two example PiRaNhA predictions; (**A**) 30S ribosomal protein S9 (PDB-ID 2J00, chain I), (**B**) CCA-adding enzyme (PDB-ID 2DRB, chain A). In the left panel, the three-dimensional structures of the protein–RNA complexes are shown, where TP, FP and FN are highlighted as cyan, yellow and pink CPK spheres, respectively, and the remaining protein residues and RNA nucleotides are represented as red and green sticks, respectively. In the right panel, the protein sequence is shown. Non-RBRs are indicated with the symbol '–' and RBRs with '+'. The TPs are highlighted in red. The prediction performance: TP (true positives), FP (false positives), TN (true negatives), FN (false negatives), Sn (sensitivity), Sp (specificity), Acc (accuracy), MCC (Mathews Correlation Coefficient) and precision are listed for each example.

recommended if potential RBRs are being selected for mutagenesis. For example, when RBR predictions are made for the RNAtestset42 data set with increasing SVM threshold levels, the precision and the specificity increase: precision and specificity at a threshold of −0.44 are 90.0 and 47.9%, respectively, at a SVM threshold of −0.20 they rise to 97.4 and 64.7% and at a threshold of 0.00 they rise to 99.1 and 76.4%.

A further point to consider is that some FP residues could in fact be true positives (i.e. be present in the RNA-binding site) in RNA–protein complexes where the PDB structure does not represent the complete functional complex. One example of this is Pop7 recently deposited in the PDB (PDB-ID 3IAB) (17). The PDB structure comprises Pop7 and Pop6 RNase MRP proteins bound to the P3 RNA domain, whereas the complete functional complex in *Saccharomyces cerevisiae* comprises more than 10 RNA secondary structure domains and many additional proteins (17). Hence, predictions made for proteins in incomplete complexes such as this, may give rise to FP residues, which are in fact novel RBRs. Such sites will only be validated when complete structures are determined.

CONCLUSION

The PiRaNhA server predicts potential RBRs based on protein sequence information alone. Four sequence-based features, such as PSSM, amino acid interface propensity, predicted residue accessibility and hydrophobicity, are integrated in a feature vector and used for predicting RBRs using an SVM model. PiRaNhA outperforms other publicly available RBR prediction servers [RNABindR (6), BindN (7) and PPRInt (9)] in a benchmark test (11). The accuracy of the PiRaNhA predictions could be improved further by the inclusion of (i) additional non-homologous RNA-binding protein structures and (ii) the structures of complete functional complexes of RNA–protein moieties into the SVM training data set. Such structures are starting to be determined as part of the many structural genomics projects (18). The aim is for the PiRaNhA server to be retrained on an updated set of non-homologous RNA-binding proteins on an annual basis, thus increasing its predictive potential.

The PiRaNhA server is of use to experimental biologists studying RNA-binding proteins in specific systems in which the structure of the protein–RNA complex is as yet unknown. The predictions made by the server allow for fewer and more targeted mutations to be made to verify RNA binding. In addition, the server is of interest to theoretical researchers wishing to analyse and compare functional residues in multiple protein data sets.

FUNDING

Strategic International Cooperative Program, Japan Science and Technology (to Y.M. and H.N.); Medical Research Council (grant number 70760 to R.V.S.).

Conflict of interest statement. None declared.

REFERENCES

1. Lukong,K.E., Chang,K.W., Khandjian,E.W. and Richard,S. (2008) RNA-binding proteins in human genetic disease. *Trends Genet.*, **24**, 416–425.
2. Jones,S., Daley,D.T., Luscombe,N.M., Berman,H.M. and Thornton,J.M. (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.
3. Ellis,J.J., Broom,M. and Jones,S. (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins*, **66**, 903–911.
4. Bahadur,R.P., Zacharias,M. and Janin,J. (2008) Dissecting protein-RNA recognition sites. *Nucleic Acids Res.*, **36**, 2705–2716.
5. Jeong,E., Chung,I.F. and Miyano,S. (2004) A neural network method for identification of RNA-interacting residues in protein. *Genome Inform.*, **15**, 105–116.
6. Terribilini,M., Sander,J.D., Lee,J.H., Zaback,P., Jernigan,R.L., Honavar,V. and Dobbs,D. (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.
7. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
8. Wang,Y., Xue,Z., Shen,G. and Xu,J. (2008) PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*, **35**, 295–302.
9. Kumar,M., Gromiha,M.M. and Raghava,G.P. (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**, 189–194.
10. Cheng,C.W., Su,E.C., Hwang,J.K., Sung,T.Y. and Hsu,W.L. (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics*, **9(Suppl. 12)**, S6.
11. Spriggs,R.V., Murakami,Y., Nakamura,H. and Jones,S. (2009) Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics*, **25**, 1492–1497.
12. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Wagner,M., Adamczak,R., Porollo,A. and Meller,J. (2005) Linear regression models for solvent accessibility prediction in proteins. *J. Comput. Biol.*, **12**, 355–369.
15. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
17. Perederina,A., Esakova,O., Quan,C., Khanova,E. and Krasilnikov,A.S. (2010) Eukaryotic ribonucleases P/MRP: the crystal structure of the P3 domain. *EMBO J.*, **29**, 761–769.
18. Nair,R., Liu,J., Soong,T.T., Acton,T.B., Everett,J.K., Kouranov,A., Fiser,A., Godzik,A., Jaroszewski,L., Orengo,C. *et al.* (2009) Structural genomics is the largest contributor of novel structural leverage. *J. Struct. Funct. Genomics*, **10**, 181–191.